Agricultural Land Markets – Efficiency and Regulation

Land value appraisal using statistical methods

Jens Kolbe, Rainer Schulz, Martin Wersing and Axel Werwatz

FORLand-Working Paper 07 (2019)

Published by
DFG Research Unit 2569 FORLand, Humboldt-Universität zu Berlin
Unter den Linden 6, D-10099 Berlin
https://www.forland.hu-berlin.de

Tel +49 (30) 2093 46845, Email gabriele.wuerth@agrar.hu-berlin.de

Land value appraisal using statistical methods

Jens Kolbe, Rainer Schulz, Martin Wersing, and Axel Werwatz*

January 18, 2019

*Kolbe and Werwatz: Technische Universität Berlin, Institut für Volkswirtschaftslehre und Wirtschaftsrecht, Straße des 17. Juni 135, 10623 Berlin, Germany, and Research Unit 2569 "Agricultural Land Markets – Efficiency and Regulation", Humboldt-Universität zu Berlin. Emails: j.kolbe@tu-berlin.de and axel.werwatz@tu-berlin.de. Schulz and Wersing: University of Aberdeen Business School, Edward Wright Building, Dunbar Street, Aberdeen AB24 3QY, United Kingdom. Emails: r.schulz@abdn.ac.uk and martin.wersing@abdn.ac.uk.

Abstract

The taxation of property based on market values requires frequent appraisals for a large number of properties. In light of the recent property tax reform discussion in Germany, it has been argued that a value-based tax therefore cannot be implemented at a reasonable cost. In several other countries, however, mass appraisal systems based on statistical methods are used for property tax assessments. In this paper, we show how this could in principle be done in Germany, using transactions data that local surveyor commissions are obliged to collect by law. We discuss the regression techniques for estimating land values from such data and illustrate them by applying them to data from Berlin, Germany. We find that the methods are capable of producing land value estimates that match up well with expert based assessments.

Keywords: land value, mass appraisal, nonparametric regression, semi-parametric regression

JEL Classification: C14, C21, H10, H20, R32, R51, R52

1 Introduction

Land values are important for real estate market participants from the private and the public sector. Examples include development companies in the private sector which need land values to decide whether projects are economically feasible. Public sector land owners need to know the market value of their land to assess the cost of alternative choices, such as market purchase versus dedicated use for housing programmes. The construction of a ring road or an airport extension may require that land is acquired through compulsory purchase and land values are needed to compensate owners for their property loss.

In all of these cases, land values are needed only for a fairly small number of properties involved. This is different for another public sector application: the taxation of property based on market values. Here, land values are required for all undeveloped sites, but will also be required for developed properties if it is taxed based on the cost value (sum of land and building value). The land value is also needed if the land and the building value of a developed property are taxed at separate graduated rates. A pure land tax, which leaves the building untaxed, is an extreme case of graduated rates. Graduated rates have theoretically appealing characteristics (Brueckner 1986, Oates and Schwab 2009), but have been implemented in only a few areas throughout the world (Bourassa 2009, Franzsen 2009).[1]

In Germany, the reform of property taxation has become an urgent matter in 2018 after the German Federal Constitutional Court (Bundesverfassungs-

[1]The windfall gains of home owners in successful cities, such as London, have led recently to renewed attention for land taxes, see: *The Economist* August 11th 2018, Leaders: Stuck in the past, 9; Briefing land-value tax: On firmer ground, 18-20.

gericht) ruled in April that the values that determine the property tax base are unconstitutional and that a solution of this problem must be found by the end of 2019. This solution, if found, must be implemented from 2025 onwards. The current market value assessments use the income and the cost approach, but rely on outdated information (from 1964 in the west part and from 1935 in the east part of Germany). While the information for assessment should have been updated every six years, this did not happen (Fuest et al. 2018). Obviously, these assessments do not reflect current relative market value gradations (Beirat BMF 2010, p.1).

The judgement of the German Federal Constitutional Court gives the legislator latitude regarding the new rules to assess the tax base, as long as it reflects the relation of properties to each other. Several proposals on the calculation of the tax base have been put forward, which include the income, the sales comparison, the cost, a pure land value, and a non-value approach that condenses the physical dimensions of a property into a numerical indicator (Fuest et al. 2018, pp.12).[2] A recurring topic in the discussion of the different proposals is the question whether a market-value tax base is feasible, given a supposed trade-off between assessment accuracy and cost (Beirat BMF 2010, p.6). Several authors see the assessment cost as argument *against* a value-based property tax. Fuest et al. (2018, pp.8), for instance, assume that a 'sufficiently accurate' assessment would cost 500EUR, which has to be spent every six years per property. Homburg (2018, p.175) is less restrained and states that value-based assessments would cost billions and give results that are only 'pseudo-accurate'. Hey (2017, p.35) is slightly more optimistic regarding the implementation of a value-based approach, although the required data

[2] McCluskey and Franzsen (2013) provide an overview of non-value approaches that are in use in other countries.

are not collected and analysed in every municipality to the same standards and with the same transparency.

Computer assisted mass appraisals have been used successfully for property tax assessment in many countries. In this paper, we draw on our own research to show how this could also be done in Germany. We discuss the statistical methods and the required data and illustrate how to combine them for land value estimation using transaction data from the city of Berlin. We are fully aware that Berlin has a very effective system of property transaction data collection and storage and that similar infrastructure is not in place in all parts of Germany. However, our research and the work of other empirical researchers shows that if such infrastructure were in place, appraisals for taxation based on statistical methods is feasible and accurate; at least when compared to land values estimates of professional appraisers. We thus arrive at a more positive conclusion: statistical mass appraisals of land values, and property-taxation schemes building upon them, could in principle be conducted in Germany at relatively low cost. We thus disagree with the statements that value based assessments are necessarily expensive and lack accuracy. We agree, however, that it seems unlikely that the infrastructure required to implement mass appraisal systems can be rolled out within the period of time set by the German Federal Constitutional Court. But mass appraisal systems cannot be blamed for this.

The rest of the paper is organised as follows. Section 2 discusses how property assessment for purposes is conducted in other countries, where we focus on those countries that use computer assisted mass appraisals. Section 3 presents methods for statistical mass appraisal for two data scenarios: (i) data from transactions of undeveloped lots and (ii) data from transactions of developed lots (i.e. single-family houses and condominiums). In each case, we illustrate the methods by applying them to transactions data in Berlin. While a discussion of statistical methods is necessarily somewhat technical, we put emphasis on what the land value estimators are actually doing. Section 4 concludes.

2 International context

In the United States, where property taxation based on current values is common, the valuation profession distinguishes between *fee appraisers*, who make individualised assessments of the market value of specific properties for business dealings, and *assessors*, who make mass appraisals for hundreds of thousands of properties. Since about 1968, and even more so today, do assessors use statistical methods, such as regression analysis, for their mass appraisals (Almy and Ferguson 2010, Back 1970). Today at least 15 countries have implemented statistical mass appraisal systems for the use in property taxation (Almy 2014, Almy and Ferguson 2010, Bidanset 2014). Among these are Australia, Canada, Egypt, Mauritius, New Zealand, South Africa, and the United States. Examples from Europe include Denmark, Finland, Latvia, Lithuania, Russia, Sweden, Northern Ireland in the United Kingdom, and the Netherlands.

Based on the experience in Canada, the Netherlands, and the United States Almy (2014) estimates that the cost per property of a high-quality statistical mass appraisal system is about 20EUR. This is in stark contrast to the 500EUR conjectured by Fuest et al, see Section 1 above, and would allow for a much more cost effective implementation of a value based property tax. Specifically, with an average tax revenue of 2,000EUR per property, the valuation cost would only be 1 percent.

The international experience shows also that statistical mass appraisals for tax purposes can provide sufficiently precise estimates of property values. Hefferan and Boyd (2010, p. 155), for example, examine the Australian experience and conclude that "[...] increasingly sophisticated computer assisted valuation techniques have, in fact, worked to assist with uniformity and consistency [of appraisals] in recent years." As Hefferan and Boyd (2010, p. 155) continue, "objection rates across Australia are well below the internationally accepted 2 per cent with many jurisdictions encountering less than 1 per cent objections." In the context of our paper, Australia is a particularly interesting example as it is one of a few countries in the world that has a land value tax *and* employs statistical mass appraisals.[3]

In Germany, it has been suggested that expert-based land values (Bodenrichtwerte, BRW) can be adopted to calculate the land value component of a reformed property tax (Beirat BMF 2010, p.2). Indeed, BRW must already be published at least every two years by independent surveyor commissions (Gutachterausschuss für Grundstückswerte, GAA). Despite being based on detailed guidelines, however, it is fair to say that BRW rely heavily on surveyors' knowledge and expertise. Given Germany's federal structure, it has thus been argued that the estimation of BRW would need to be harmonized in order to assure a uniform quality standard for tax purposes (Löhr 2011).

We discuss next how statistical methods can assist the mass appraisal of land values in a transparent manner.[4]

[3]Lithuania is an European example for such a country (Baranska 2013).

[4]McCluskey et al. (2013) provide an overview of how statistical mass appraisal systems can be implemented for the use in property taxation.

3 Statistical methods and results

We begin our discussion of the methodology with two figures illustrating what we want to learn from the data and what the data is assumed to consist of (at the bare minimum).

Figure 1 shows a land value map for Berlin based on the BRW values published by Berlin's GAA. The land values are for the year 2009, as they are based on information up to and including that year. Land values within the first percentile of the value distribution are shown in light grey (bright yellow). The intensity increases to dark grey (bright red) for land values that fall within the tenth percentile of the distribution. From this distant view, the dominating feature of the map is the declining color intensity in the outward direction away from the city centre.[5]

For the present purpose, the central aspect of the map is that it delivers an estimate of the expected price of land at any desired location in Berlin. We will denote such an estimate as $\hat{\theta}_{BRW}\left(l_1, l_2\right)$, where l_1 and l_2 are the latitude and longitude of the location, respectively.

The second figure, Figure 2, shows the locations of 24,519 arms-length transactions of undeveloped land that occurred in Berlin during 1996-2009. The transaction data comes also from Berlin's GAA. Most transactions of undeveloped land took place in the residential areas at the outskirts of Berlin. Less transactions of undeveloped land took place in the densely developed city centre. The expert-based values shown in Figure 1 are based on this data, but the GAA surveyors will have considered also other real estate market information. In the following section, we will consider how to estimate a land

[5]At the aggregate level, the land value map thus confirms the prediction of the monocentric city model that the land rent gradient falls with distance from the city centre.

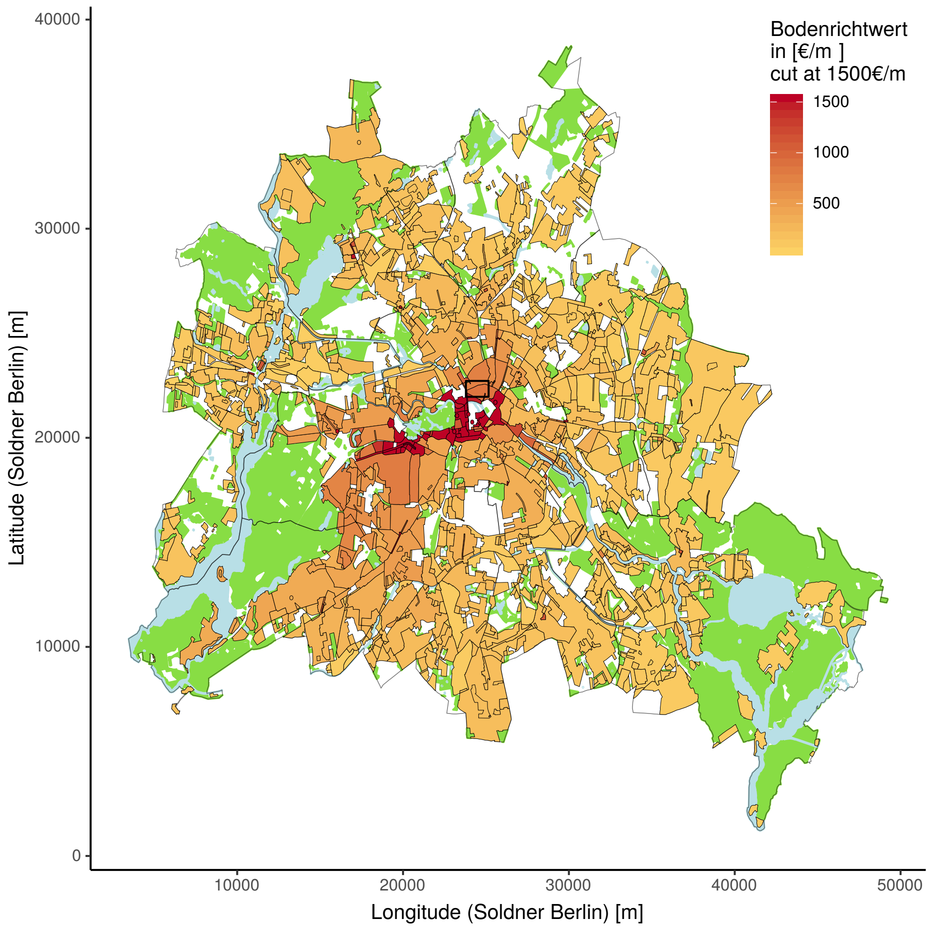

Figure 1: Expert-based land value (BRW) map for Berlin. Shows map of expert-based land values (BRW, in logs) for Berlin. Reference date is 1 January 2010. Source: Geoportal Berlin/Bodenrichtwerte 01.01.2010.

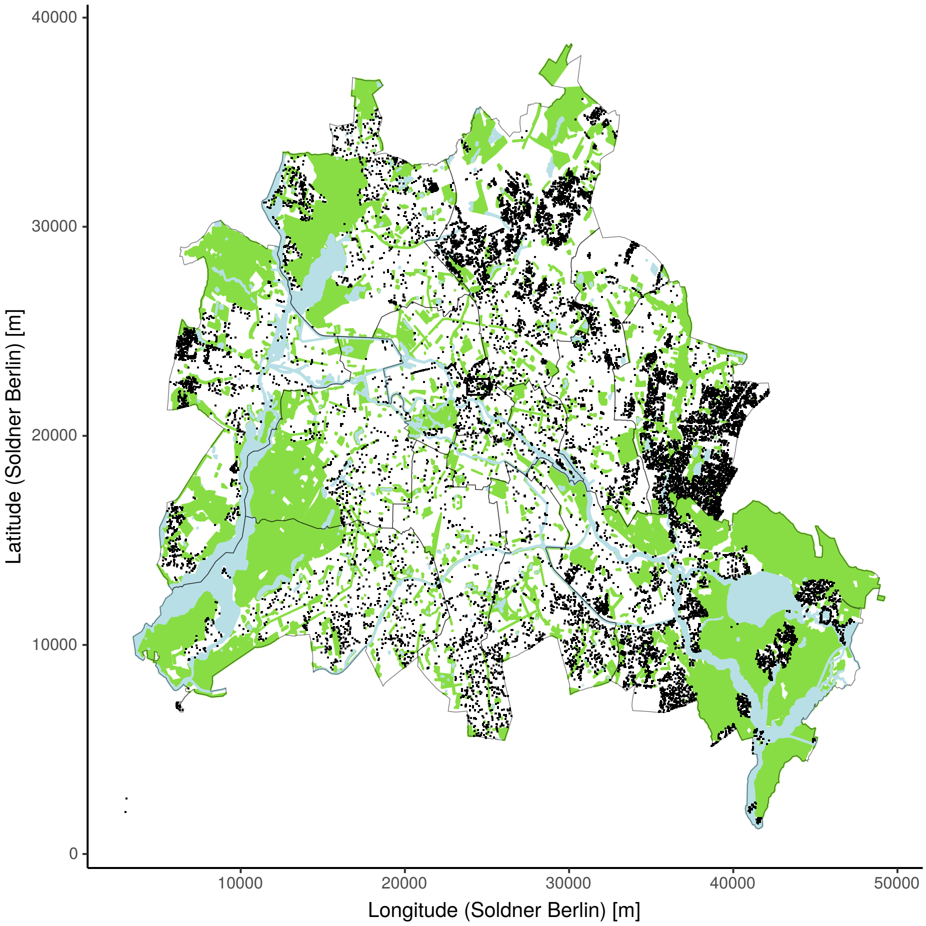

Figure 2: Location of transacted sites within Berlin. Shows the location of 24,519 undeveloped sites that have been transacted between 1996 to 2009. Solid lines represent the borders of Berlin's 12 administrative districts (as of the year 2000).

value surface like that in Figure 1 from data such as that shown in Figure 2 *alone*. That is, we will consider the problem of estimating the expected land value at a given location from geocoded data of transactions of undeveloped land. Sales of undeveloped land directly contain information about the value of land. However, as seen in Figure 2, undeveloped land is typically not available and sold in the city centre. In this most valuable area of a city, often only sales of developed properties are found. Observed market prices of such properties contain information about the value of the underlying land *and* the structure erected on it. The statistical analysis aiming to extract land values from transactions of developed properties must thus find a way to separate the land and building value components. How this can be done is the topic of Subsection 3.3 below.

3.1 Statistical land value estimation with nonparametric kernel regression

Required data input. In the following section, we assume that the transactions data has information on (a) the price and (b) the location of a sale only. In short, the data is assumed to consist of n observations of the form

$$\{p_i, l_{1i}, l_{2i}\}, i = 1, \ldots, n$$

where $p_i = \ln(P_i)$ is the log land price per square meter of the lot and l_{1i} and l_{2i} are its latitude and longitude. Why researchers typically work with *log* prices, rather than the prices themselves is explained below.

The regression model. From a statistical point of view, a land value map such as that in Figure 1 is viewed as an estimate of a regression model

$$p_i = \theta(l_{1i}, l_{2i}) + \epsilon_i \ , \tag{1}$$

where $\theta(l_{1i}, l_{2i})$ is the expected (log) land value at location l_{1i}, l_{2i}. From an applied perspective, $\theta(l_{1i}, l_{2i})$ is the aim of land value appraisal at a location. The error term ϵ_i captures deviations of the log price of a specific lot (lot i) from its expected value. This transaction noise is assumed to average out at any given location and to have the same amount of variation at all locations. The latter assumption is not needed for the regression based estimation of land values, but is typically invoked when error bounds around the land value estimate are also desired. It tends to be approximately satisfied for log prices but violated for the prices themselves, thus providing a reason why the researchers usually uses log prices as the dependent variable.

The estimation method. As the locations l_1, l_2 are varied, a surface of expected land values at various locations arises which graphically can be represented by a coloured map such as Figure 1. It is the "regression surface" in statistical terms. A standard method to estimate such a surface at any desired location from data such as that depicted in Figure 2 is kernel regression. It merely assumes that the regression surface has no jumps. In terms of the present context: that land values change smoothly in space and not abruptly. Apart from the smoothness requirement, the form of the estimated land value map is not restricted a priori. In particular, no specific formula is imposed on the data in this "nonparametric" procedure. It can thus freely adapt to the information in the data of how log prices vary from location to location. It does so by forming local averages of log prices.

The kernel regression estimator introduced by Nadaraya (1964) and Watson (1964) is formally defined as

$$\widehat{\theta}_{NKR}(l_1, l_2) = \sum_{i=1}^{n} \frac{K\left(\frac{l_1 - l_{i1}}{h_1}, \frac{l_2 - l_{i2}}{h_2}\right)}{\sum_{j=1}^{n} K\left(\frac{l_1 - l_{i1}}{h_1}, \frac{l_2 - l_{i2}}{h_2}\right)} \cdot p_i \tag{2}$$

or, more briefly

$$\widehat{\theta}_{NKR}(l_1, l_2) = \sum_{i=1}^{n} W_{i,h_1,h_2}(l_1, l_2) \cdot p_i \tag{3}$$

The second version, Eq. 3, highlights that the kernel regression estimator indeed can be seen as a weighted (local) average of the the log price p_i with weights $W_{\bullet,h_1,h_2}(\bullet)$. The first version, Eq. 2, shows more explicitly how these weights are formed from kernel functions $K(\bullet)$, that give the method its name.

How transactions are weighted. Various specific formulas have been proposed for the kernel function $K(\bullet,\bullet)$. A popular choice is to specify $K(\bullet,\bullet)$ as the product of two separate kernel functions that work on one distance only,

$$K\left(\frac{l_1 - l_{i1}}{h_1}, \frac{l_2 - l_{i2}}{h_2}\right) = K_1\left(\frac{l_1 - l_{i1}}{h_1}\right) K_2\left(\frac{l_2 - l_{i2}}{h_2}\right) \tag{4}$$

and to use for both for both component kernel functions the formula proposed by Epanechnikov (1969):

$$K_j\left(\frac{l_j - l_{ij}}{h_j}\right) = \underbrace{I\left(\left|\frac{l_j - l_{ij}}{h_j}\right| \leq 1\right)}_{\text{part 1}} \; \underbrace{\frac{3}{4}\left\{1 - \left(\frac{l_j - l_{ij}}{h_j}\right)^2\right\}}_{\text{part 2}} \tag{5}$$

This weighting function has two parts. The first part is an indicator function that sets the weight of an observation outside the target location's vicinity (in direction l_j) to 0. The second part assigns weights to the observations inside the vicinity according to the graph shown in Figure 3.

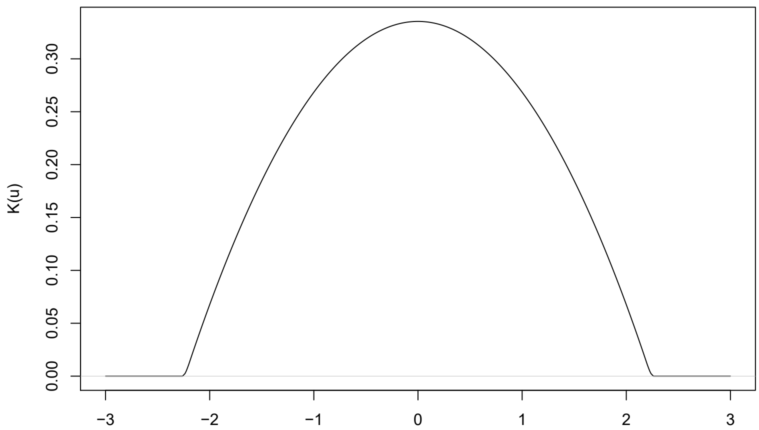

Figure 3: Epanechnikov kernel function. X-axis represents relative distance from the target location. Y-axis represents the kernel weight from Eq. 5. Bandwidth is set to $h_j = 1$.

Clearly, maximum weight is given if the relative distance is zero and weights decline in the depicted pattern with increasing distance to the target location.

In Figure 4 we illustrate the two-dimensional weighting with latitude and longitude axes in the map format.

Only data points within the rectangle that is defined by the bandwidths and is surrounding the target location, will receive a nonzero weight in the local averaging procedure. Here, coordinates with equal weight are represented by contour lines. Contour lines closer to the target location (red dot) are narrower and represent a higher weight.

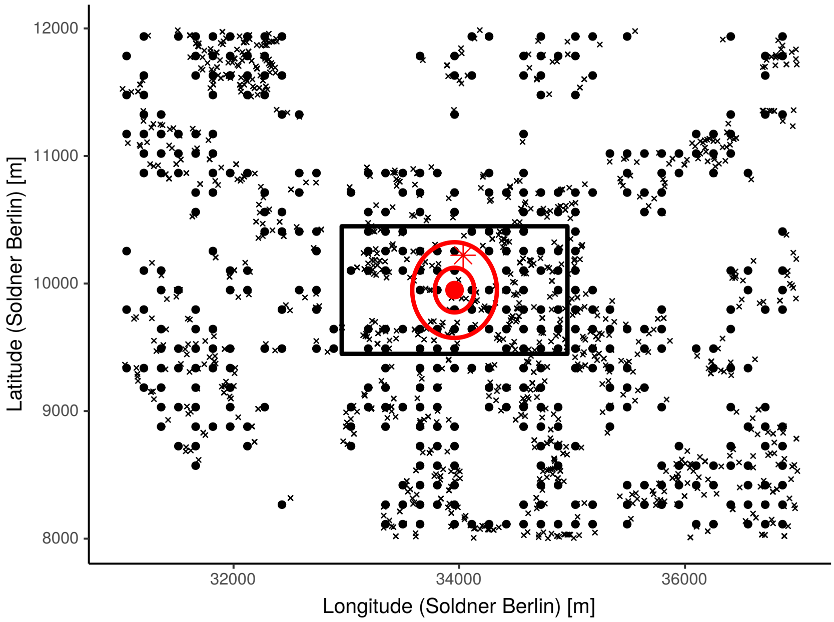

Figure 4: Illustration of nonparametric kernel regression. Shows grid points (dots) and observation (crosses) for a neighborhood in Treptow-Köpenick. The red dot represents the target location. Kernel weights are represented by (red) contour lines. Each observed transaction on a contour line receives the same weight when the estimating the land value of the target location.

Infobox 1: Summary of kernel regression procedure

- Kernel regression works at a specific desired location. It is thus a local procedure, just as land value assessment always works locally.
- The desired location may or may not be in the data.
- At the desired location, kernel regression simply forms a weighted sample average over the log prices of lots sold in the vicinity.
- The vicinity is specified by an ellipsoid around the location. The length and width of the ellipsoid is controlled by the 'bandwidths', that must be specified by the user.

- The average considers each observation but those observations outside the ellipsoid will receive zero weight. How much weight the (log) price of an observation receives is determined by the kernel function.
- For determining each weight, the kernel function considers the distances of an observation from the location at which the estimate is desired in both the longitude and latitude directions. These distances are computed relative to the bandwidth. For each observation inside the vicinity rectangle, both relative distances are smaller or equal to 1 in absolute value.
- The closer an observation is to the target location, the more weight it will receive. Maximum weight is thus given to a lot that is exactly located where an estimate of the expected log price is required.

Sample data. We illustrate the procedure with the transaction data from Figure 2. In Germany, GAAs are entitled by law to request and collect information on all real estate transactions. Their data bases thus provide a rich source for the regression based land value estimation described here. Our data from Berlin provides for each observation the required input: the (log) transaction price per square meter (sqm) and geocoordinates. A detailed data description can be found in Kolbe et al. (2015).

Regression based land value estimates. We applied the kernel regression estimator of Eq. 3 to estimate land value on a grid of other locations. This allows us to produce a land value map for Berlin.[6] To get the land value

[6]In a similar fashion, McMillen (1996) estimates a land value surface for Chicago via kernel regressions.

estimates on the “natural” scale, we re-transform the estimated log land values to EUR using the formula from Kennedy (1983).

The estimated land value map is shown in Figure 5. For coloring, we employ the same scheme as the BRW map of Figure 1 above.

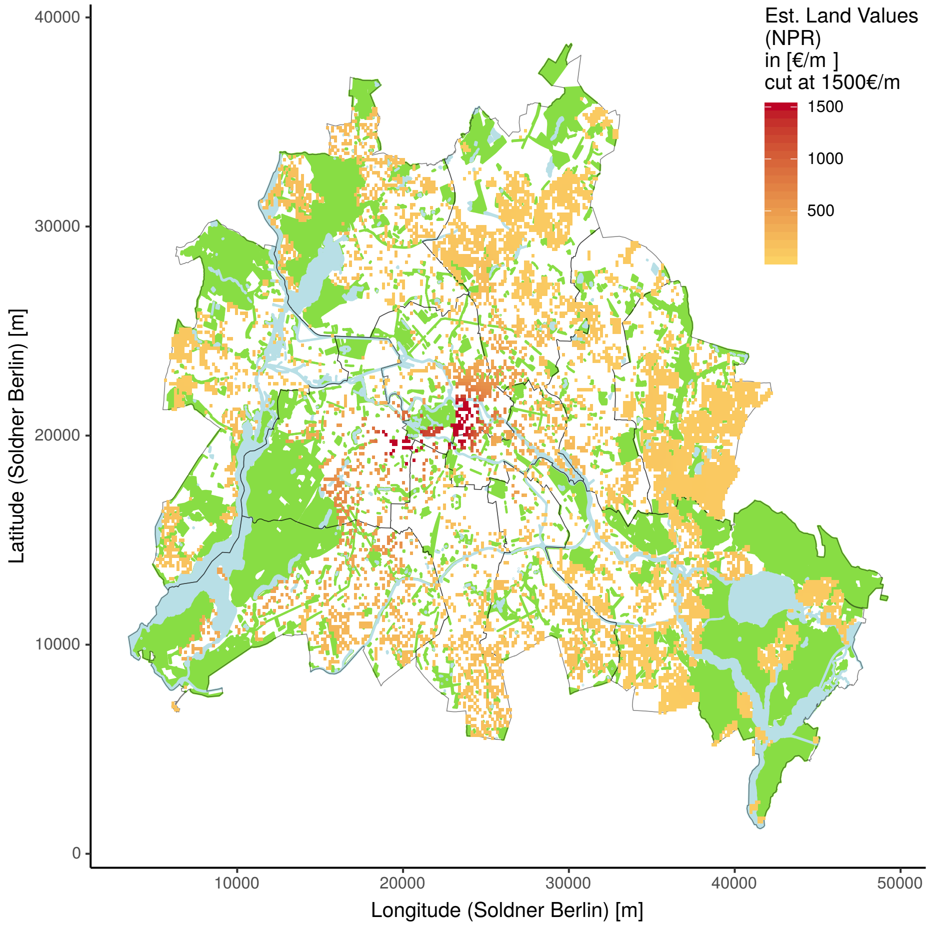

Figure 5: Kernel regressionbased land value map for Berlin. Shows contour map of estimated land values using kernel regression. Bandwidth were set to $h_1 = h_2 = 650$.

Where possible, a comparison of both land value maps shows that colors (and thus estimated land values) largely agree. The coefficient of correlation between the two land values is 0.704. The kernel regression estimates are solely based on the transactions data data of Figure 2. They could thus only be computed away from the city centre where sales of undeveloped land occurred. The BRW map of Figure 1, on the other hand, covers the entire Berlin area. The GAA surveyors must have considered other information than just transactions on undeveloped land in order to arrive at their comprehensive set of land value estimates. In particular, they must have incorporated information about how market participants valued land in the city centre. In this area, virtually all land is developed. Below, we will thus consider how to do regression based estimation of land values from transactions data of developed properties.

Bandwidth selection As demonstrated, kernel regression produces a local estimate of the expected land value by averaging over the log prices of observations in the vicinity of the target location. A key question then is how large the size of this vicinity should be. It is determined by the bandwidths $h_1 = 2,000$ and $h_2 = 1,000$ metres, but provided no justification for this particular choice.

A first impulse may suggest to make h_1 and h_2 as small as possible, i.e. to set them to 0. This would ensure that only those transactions are included in the average that occurred exactly at the target location for which a land value estimate is desired. This would prevent any bias in the land value estimate that may arise if near-by transactions sell for a (slightly) higher or lower price. However, there may be very few or even no observations at the target location making estimation very unreliable or altogether infeasible. From this perspective, larger bandwidths are desirable as they ensure that more observations enter the calculation. An optimal choice of the bandwidths strikes a balance between these two concerns.

It has been demonstrated that such bandwidth values can be found in a data-driven, objective way by the method of cross-validation. As the name suggests, cross-validation uses the data to "validate" the estimates and compute an overall performance measure for a given set of bandwidths. It then repeats this process for many different set of bandwidth and finally selects those two bandwidth values that deliver the best "validated" performance. This procedure ensures that the land value estimate includes enough observations to be reasonably reliable but still gets the local level of the average land value approximately right.

Statistical properties. All valuation methods are only *estimates* of the true land value at a certain location. It is a major advantage of the statistical approach described here that it allows to quantify the uncertainty that is an unavoidable aspect of all estimation methods. Since the kernel method is essentially a weighted local average of log prices it is not difficult to estimate its precision. It is given by the ratio of the variance (a measure of the "noise" in log prices) and the effective size of the sample that entered the average. Details are given in Härdle (1990, Ch. 4). This can be used to construct confidence intervals around the land value estimate that cover the true land value with a prespecified confidence level.

Related literature and alternative methods. In addition to kernel regression there are other nonparametric estimation methods that are also based on local averaging.[7] Colwell and Munneke (2003), for instance, estimate location values for Chicago from transactions of undeveloped land using smoothing splines. Just as kernel regressions, smoothing splines assume also that land values may not change abruptly from location to location. This assumption,

[7] Härdle et al. (2004, Ch. 4) provide an overview of alternative nonparametric methods.

however, is at odds with the blockwise outlay of cities where adjacent neighborhoods can be sharply demarcated by roads and may distinctly differ in their character. Indeed, the detail of the BRW map published by Berlin's surveyor commission shown in Figure 6 reveals that the land value surface produced by these experts is not smooth.

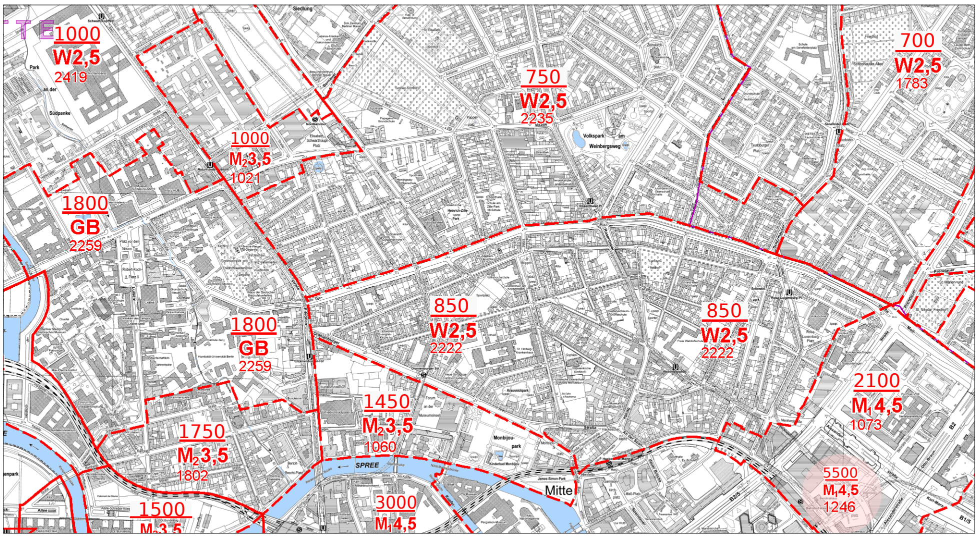

Figure 6: Detail of expert-based land value (BRW) map. Shows central business district including the boulevard Unter den Linden, the Museumsinsel, and the Alexanderplatz. Reference date for map is 1 January 2010. Source: Geoportal Berlin/Bodenrichtwerte 01.01.2010.

3.2 Identifying land value zones with nonparametric adaptive regression

In Kolbe et al. (2015) we use Adaptive Weights Smoothing (AWS) to estimate a piecewise constant land value map akin to the map published by the Berlin's GAA in their *Bodenrichtwertatlas* (BRW Atlas). Adaptive Weights Smoothing directly builds on the kernel regression estimator with a slightly simplified kernel function. Specifically, the kernel of Eq. 4 becomes

$$(6)\quad K\left(\frac{l_{1i}-l_{1j}}{h_1},\frac{l_{2i}-l_{2j}}{h_2}\right)=K\left(\frac{|l_{1i}-l_{1j}|+|l_{2i}-l_{2j}|}{h}\right)=K\left(distance^1_{ij}\right)\ .$$

That is, distance is now simply measured as the sum of the absolute latitude and longitude distances, divided by a common bandwidth. This bandwidth is set to a small number to obtain the initial land value estimate, denoted as $\hat{\theta}^0(l_{1i}, l_{2i})$, or $\hat{\theta}^0_i$ in short. The 0 superscript denotes the initial start-up iteration. That is, only few spatially close observations are used to form the resulting initial land value estimate $\hat{\theta}^0_i$ at any location (l_{1i}, l_{2i}).

In the next step (and all subsequent steps), the land value estimate is still a local weighted average.

$$(7)\qquad \hat{\theta}^1(l_{1i}, l_{2i})=\frac{\sum_{j=1}^n w^1_{ij}p_j}{\sum_{j=1}^n w^1_{ij}}$$

The key difference is that a second kernel function enters the weighting of observations. The weight of an observation is now formed by a product of the familiar “distance kernel” of Eq.6 and a “level kernel”: $w^1_{ij}=K\left(distance^1_{ij}\right)\times K\left(level^1_{ij}\right)$, where

$$(8)\qquad level^1_{ij}=\left(\frac{\hat{\theta}^0_i-\hat{\theta}^0_j}{\sqrt{2\hat{\sigma}^2_\epsilon}}\right)^2\cdot\frac{n^0_i}{\lambda}$$

Here, the key argument is $\hat{\theta}^0_i-\hat{\theta}^0_j$, which represents the difference between the land value estimates at location l_{1j}, l_{2j} and l_{1i}, l_{2i} in iteration 0. If the two estimates are quite close, both locations appear to have rather similar land values. Consequently, the observed (log) land price from location l_{1j}, l_{2j} receives substantial weight to form a land value estimate at location l_{1i}, l_{2i} in the subsequent iteration. It is this feature that makes AWS ‘structurally adaptive’ and allows to find data-driven areas areas of homogenous land value alike the

Bodenrichtwertzonen defined by the GAA. This is achieved by successively increasing the latitude/longitude bandwidth in successive iterations and using the smoothing parameter λ in Eq. 8 as a threshold for judging the closeness of land value estimates from different locations.

Estimated land values. Because AWS is a local, iterative procedure it is computationally intensive. We thus computed land value estimates not at individual coordinates but on a grid of "bins", that is small squares of size 152x152 metres. The estimated land values for each bin are shown inFigure 7, where we use the same coloring as in the BRW map of Figure 1.

Comparing both maps shows that, just as the kernel estimates, the AWS bins in Figure 7 only cover a part of the continuously shaded BRW areas of Figure 1. Obviously, the expert-based BRW map must have used other information on land value where no sales of undeveloped were available. Where comparisons can be made, AWS and BRW land value estimates agree fairly well in terms of their spatial color patterns.

Smoothing parameter selection and validation with BRW Like any local, nonparametric statistical method the performance of AWS depends on a smoothing parameter that must be specified by the user. In the case of AWS, the parameter λ in Eq. 8 determines how responsive the estimator is to local variation in the estimated land value level in the previous iteration. How to optimally choose λ is the topic of ongoing research. The estimates in Figure 7 were computed using the value $\lambda^* = 19.9$ suggested by Polzehl and Spokoiny (2006, 2008).

In addition, we ran AWS for the eight different values reported in the top row of Table 1. We found the performance, as measured by the coefficient of

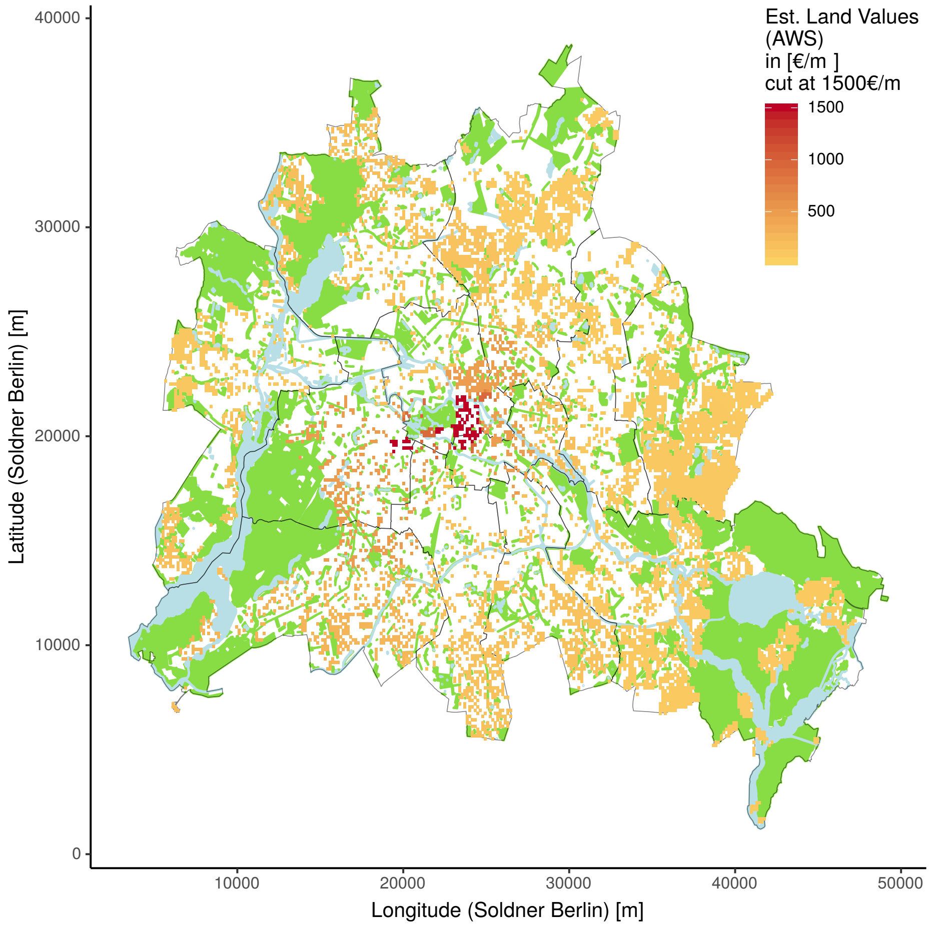

Figure 7: AWS-based land value map for Berlin. Shows contour map of estimated AWS land values. Level bandwidth is set to $\lambda^* = 19.9$.

determination R^2, of the procedure to be rather insensitive to which value of λ is employed.

Table 1: Explanatory power. Reports coefficient of determination R^2 for bivariate regressions of BRW_i and land prices y_i on AWS land values $\widehat{\theta}_i$. Regressions include a constant. Number of observations used for regressions in first row is 7,222 and 7,448 for regressions in second row.

	λ^*	$\boldsymbol{\lambda}$						
	19.9	3.8415	4.4756	10.5180	16.8410	23.2840	29.7938	36.346
BRW	0.7747	0.7274	0.7390	0.7640	0.7720	0.7828	0.7733	0.7690
Land price	0.6992	0.8661	0.8526	0.7734	0.7195	0.6764	0.6525	0.6418

The R^2 values were obtaind by regressing the AWS land values obtained for a given value of λ on either observed land prices or BRW values. Apparently, the agreement between AWS land value estimates and, both, land prices and BRW land values is fairly good for all values of λ. In a more detailed analysis reported in Kolbe et al. (2015), we found that AWS also determines similar areas of homogenous land values as the BRW Atlas. In summary, we found AWS to be a transparent statistical procedure capable of estimating land values close to the expert benchmark, both, in their level as well as in their geographical structure.

Related literature. AWS has also been applied by Helbing et al. (2017) to estimate agricultural land values. An alternative to AWS for nonparametric regression, when the regression surface may have jumps or edges, is the wavelet method described, for instance, in Vidakovic (1999).

3.3 Estimating land values from property transactions using Semiparametric Regression of

Both kernel regression and Adaptive Weights Smoothing estimate land values by averaging over prices of undeveloped lots and thus work with the most direct and "clean" market information on the value of land at a given location. However, such information tends to be unavailable in the centre of a city where virtually all lots are developed. Data on transactions data of developed lots, thought, contains information about the bundle of land *and* building. Hence, for land value estimation, observed prices of houses or condominiums need to be split into their land and building component. This can also be achieved by a regression analysis.

Required data input. In this section, we assume that the transactions data has information on (a) the price, (b) the location and (c) building characteristics of a property. In short, the data is assumed to consist of n observations of the form

$$\{p_i, l_{1i}, l_{2i}, X_{1i}, \ldots, X_{pi}\}, i = 1, \ldots, n$$

where $p_i = \ln(P_i)$ is the log land price of the property, l_{1i} and l_{2i} are its latitude and longitude and $X_{1i}, \ldots, X_{pi}$ are building characteristics such as floor space or age.

Semiparametric regression model. We start with the assumption that the log price of a property can be split into the value of the building and the value of land to obtain the partial-linear regression model

$$p_i = \beta_0 + \underbrace{\beta_1 X_{1i} + \ldots + \beta_p X_{pi}}_{\text{building}} + \underbrace{\theta(l_{1i}, l_{2i})}_{\text{land}} + \epsilon_i \ . \tag{9}$$

Here, the coefficients $\beta_1, \ldots, \beta_p$ quantify the influence on the associated building characteristics on the log price of the property and all variables are measured per square meter lot size.

Compared to our initial regression model in Eq. 1, the model of Eq. 9 has added the building part that linearly combines variables and their coefficients ("parameters"). It thus additively combines a parametric building component with a nonparametric landcomponent, making the model "semiparametric" in statistical parlance.

Two-step estimation. Our goal remains to estimate the nonparametric land value surface $\theta(l_{1i}, l_{2i})$. To do so, we estimate in a first step the coefficients of the building part. Denote these coefficient estimates as $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. We use them for removing the building value from the property price

$$\hat{u}_i \quad = \quad p_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_p X_{pi}) \ . \tag{10}$$

The resulting deviations of property price from building value, the "residuals" $\hat{u}_i$, are used as the dependent variable in the second step. In this final step, the residuals $\hat{u}_i$ are regressed on the location coordinates l_{1i} and l_{2i} using a nonparametric regression procedure such as kernel regression or AWS. We illustrate this two-step approach again with transaction data from Berlin.

Sample data. We illustrate the procedure with Berlin data combining 27,549 single-family house transactions with 166,839 sales of condominiums that occurred between 1996 and 2013. Condominium transactions provide information about land values in the centre of Berlin whereas single-family house transactions predominantly happen in the outskirts. Hence, it is necessary to include sales information for both types of properties to obtain a comprehensive set of land value estimates. The flip side is that the model in Eq. 9 needs to have two sets of regression coefficients, one for each type of dwelling, to accommodate their different specifications.

Semiparametric land value estimates for Berlin. To estimate the coefficients of the building characteristics in the first step, we employed the estimator proposed by Yatchew (1997). The basic idea of the estimator is that the land value $\theta(l_1, l_2)$ can be neglected when considering differences of prices of near-by observations. Hence, the data are ordered to be geographically close to each other and then differences in prices are regressed on differences in building characteristics by ordinary least squares to estimate the building component coefficients. In the second step, we employed AWS to estimate the land value surface shown in Figure 8.

In Kolbe et al. (2012), we calculated the correlation between the BRW values and our semiparametric land value estimates based on house transactions only. For this subset of the data, we found a strong positive correlation of 0.845, indicating that the two-step semiparametric regression approach described in this section is capable of extracting valid land values from data on property transactions.

Related literature. Only a few previous studies have modeled location values from house price information using semiparametric regressions. Cheshire and Sheppard (1995) and Bryan and Sarte (2009) are examples; none of these studies compares the estimated land values with benchmarks.[8] Anglin and

[8]Lack of such a benchmark is the reason why land values have to be estimated in the first place.

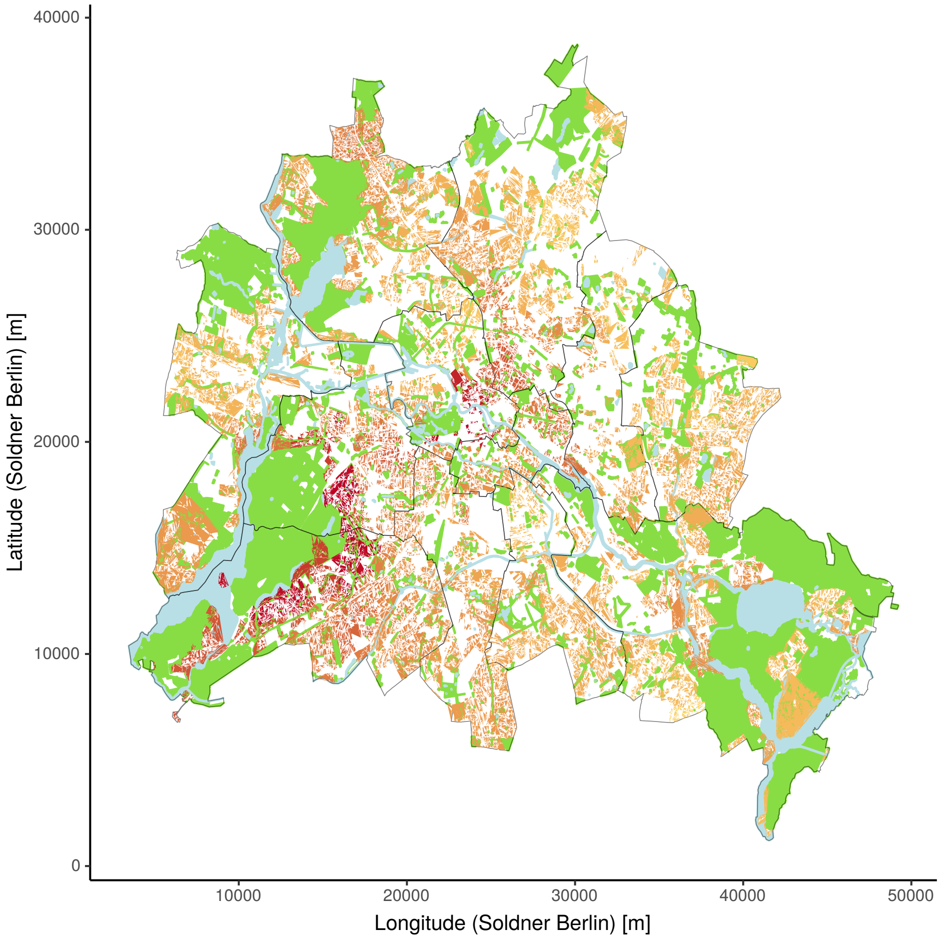

Figure 8: Land value map based on semiparametric regression. Shows contour map of land values estimated from the residuals in Eq. 10 via AWS. Level bandwidth is set to $\lambda^* = 19.9$.

Gencay (1996), Clapp (2003) and Case et al. (2004), among others, fitted semiparametric regression models to house prices in order to evaluate their suitability for mass appraisals. The results of these studies show that the nonparametric modeling of the location component improves the accuracy of house price appraisals relative to more standard methods, such as parametric linear regression.

4 Conclusion

In several other countries, computer assisted mass appraisal systems are used for property tax assessments, but this topic has not featured prominently in the reform discussion in Germany.[9] In this paper, we draw on our own research to show how this could also be done in Germany using the transactions data that local surveyor commissions are obliged to collect by law. We have considered two types of transactions data: geocoded sales of undeveloped land and geocoded sales of developed properties. The former present the purest source of land valuation by market participants. The latter need a statistical "separation" of the sales price into a building component and a land component. In both cases, we have demonstrated how to use modern nonparametric regression techniques to estimate land values at a given location where sales occurred in the vicinity. We have found in our application of these methods to data from Berlin that the resulting land value estimates typically agree well with (more expensive) expert based land values. Unlike those expert based estimates, the statistical methods we used are transparent and can be standardised.

We are aware that there are several areas that need further investigation. First, there might be municipalities that are characterised by fairly low num-

[9]Exceptions are Senatorin für Finanzen (2010) and Houben (2017).

bers of transactions. Additional information such as list prices might be a useful data extension. The volume of list prices is always much higher than the number of actual transactions, simply, because not every listing leads to a transaction. This requires an examination whether list prices can complement transaction prices. Second, there is the question about the cost and the efficient organisation of the administrative process. There is evidence that property assessment for tax purposes has the potential to be conducted with economies of scale (Sjoquist and Walker 1999) and that assessment offices can have an optimal size (Krupa 2017).[10] It would also be very interesting to assess the relationship between cost and accuracy in a rigourous manner as suggested by Mehta and Giertz (1996). Finally, our paper focussed on market values of undeveloped land, but statistical methods can be used equivalently for the assessment of developed land, see for instance Schulz et al. (2014).

Acknowledgement

Kolbe and Werwatz thank the Deutsche Forschungsgemeinschaft, DFG research unit FOR2569 “Agricultural Land Markets – Efficiency and Regulation” for financial support. The usual disclaimer applies.

[10]Neither of the two studies takes explicit account of tax administrations that use computer assisted mass appraisal systems, although Sjoquist and Walker (1999) mention that several of the offices in their sample use them.

References

Almy, R. R.: 2014, Valuation assessment of immovable property, *Working Papers on Fiscal Federalism 19*, OECD.

Almy, R. R. and Ferguson, A. G.: 2010, Valuing our world: Potential roles for AVMs and CAMA, *Proceedings of the Union of Pan-American Valuers' Association XXV Congress*, Appraisal Institute, Chicago.
URL: *http://www.mrcl.com.br/upav/19.pdf*

Anglin, P. M. and Gencay, R.: 1996, Semiparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **11**, 633–648.

Back, K.: 1970, Land value taxation in light of current assessment theory and practice, *in* D. M. Holland (ed.), *Land value taxation in light of current assessment theory and practice*, number 5 in *Committee on Taxation, Resources and Economic Development*, University of Wisconsin Press, pp. 37–54.

Baranska, A.: 2013, Real estate mass appraisals in selected countries: Functioning systems and proposed solutions, *Real Estate Management and Valuation* **21**, 35–42.

Beirat BMF, W.: 2010, Reform der Grundsteuer, *Stellungnahme*, Bundesministerium der Finanzen, Berlin.

Bidanset, P. E.: 2014, Moving automated valuation models out of the box: The global geography of AVMs, *Fair and Equitable* pp. 3–7.
URL: *https://www.iaao.org/media/Topics/AVMs/FE_July_Bidanset.pdf*

Bourassa, S. C.: 2009, The U.S. experience, *in* R. F. Dye and R. W. England (eds), *Land value taxation. Theory, evidence, and practice*, Lincoln Institute of Land Policy, Cambridge MA, pp. 11–26.

Brueckner, J. K.: 1986, A modern analysis of the effects of site value taxation, *National Tax Journal* **39**, 49–58.

Bryan, K. A. and Sarte, P.-D. G.: 2009, Semiparametric estimation of land price gradients using large data sets, *Economic Quarterly* **95**, 53–74.

Case, B., Clapp, J. M., Durbin, R. and Rodriguez, M.: 2004, Modeling spatial and temporal house price patterns: A comparison of four models, *Journal of Real Estate Finance and Economics* **29**, 167–191.

Cheshire, P. and Sheppard, S.: 1995, On the price of land and the value of amenities, *Economica* **62**, 247–267.

Clapp, J. M.: 2003, A semiparametric method for valueing residential locations: Applications to automated valuation, *Journal of Real Estate Finance and Economics* **27**, 303–320.

Colwell, P. F. and Munneke, H. J.: 2003, Estimating a price surface for vacant land in an urban area, *Land Economics* **79**, 15–28.

Epanechnikov, V. A.: 1969, Non-parametric estimation of a multivariate probability density read more: https://epubs.siam.org/doi/10.1137/1114019, *Theory of Probability & Its Applications* **14**, 153–158.

Franzsen, R. C. D.: 2009, International experience, *in* R. F. Dye and R. W. England (eds), *Land value taxation. Theory, evidence, and practice*, Lincoln Institute of Land Policy, Cambridge MA, pp. 27–50.

Fuest, C., Immel, L., Meier, V. and Neumeier, F.: 2018, Die Grundsteuer in Deutschland: Finanzwissenschaftliche Analyse und Reformoptionen, *ifo*

Studie, Forschungsgruppe Steuer- und Finanzpolitik, Leibnitz-Institut für Wirtschaftsforschung an der Universität München e. V. Studie im Auftrag von Haus & Grund Deutschland–Zentralverband der Deutschen Haus-, Wohnungs und Grundeigentümer e.V. sowie ZIA Zentraler Immobilien Ausschuss e.V.

Härdle, W. K.: 1990, *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press, Cambridge.

Härdle, W. K., Müller, M., Sperlich, S. and Werwatz, A.: 2004, *Nonparametric and Semiparametric Models*, Springer-Verlag, Berlin.

Hefferan, M. J. and Boyd, T.: 2010, Property taxation and mass appraisal valuations in Australia and New Zealand, *Property Management* **28**, 149–162.

Helbing, G., Shen, Z., Odening, M. and Ritter, M.: 2017, Estimating location values of agricultural land, *The German Journal of Agricultural Economics* **66**, 188–201.

Hey, J.: 2017, Verfassungsmäßigkeit der Reform der Bemessungsgrundlage der Grundsteuer und der Entwicklung der Grundsteuerhebesätze, *Gutachten im Auftrag der BID Bundesarbeitsgemeinschaft Immobilienwirtschaft Deutschland*, Institut für Steuerrecht Universität zu Köln.

Homburg, S.: 2018, Ein Vorschlag zur Grundsteuerreform, *Wirtschaftsdienst* **98**, 169–175.

Houben, H.: 2017, Bewertung für grundsteuerliche Zwecke zwischen Wunsch und Wirklichkeit, *Steuer und Wirtschaft* (2), 184–199.

Kennedy, P.: 1983, Logarithmic dependent variables in prediction bias, *Oxford Bulletin of Economics and Statistics* **45**, 389–392.

Kolbe, J., Schulz, R., Wersing, M. and Werwatz, A.: 2012, Location, location, location: Extracting location value from house prices, *SFB 649 Discussion Paper* No 2012-040.
URL: *http://SFB649.WIWI.HU-BERLIN.DE/PAPERS/PDF/SFB649DP2012-040.PDF*

Kolbe, J., Schulz, R., Wersing, M. and WerwatzI, A.: 2015, Identifying Berlin's land value map using adaptive weights smoothing, *Computational Statistics* **30**, 767–790.

Krupa, O.: 2017, Government consolidation in property tax administration, *State and Local Government Review* **49**, 27–36.

Löhr, D.: 2011, Reform der Grundsteuer: Zu einem blinden Fleck in der Stellungname des Wissenschaftlichen Beirats beim Bundesministerium der Finanzen, *Wirtschaftsdienst* **91**, 333–338.

McCluskey, W. J., Davis, P., McCord, M., McIlhatton, D. and Haran, M.: 2013, Computer assisted mass appraisal and the property tax, *in* W. J. McCluskey, G. C. Cornia and L. C. Walters (eds), *A Primer on Property Tax. Administration and Policy*, Wiley-Blackwell, Chichester, chapter 14, pp. 307–338.

McCluskey, W. J. and Franzsen, R.: 2013, Non-market value and hybrid approches to property taxation, *in* W. J. McCluskey, G. C. Cornia and L. C. Walters (eds), *A Primer on Property Tax. Administration and Policy*, Wiley-Blackwell, Chichester, chapter 13, pp. 287–305.

McMillen, D. P.: 1996, One hundred fifty years of land values in Chicago: A nonparametric approach, *Journal of Urban Economics* **40**, 100–124.

Mehta, S. and Giertz, F.: 1996, Measurering the performance of the property tax assessment process, *National Tax Journal* **49**, 73–85.

Nadaraya, E. A.: 1964, On estimating regression, *Theory of Probability & Its Applications* **9**, 141–142.

Oates, W. E. and Schwab, R. M.: 2009, The simple analytics of land value taxation, *in* R. F. Dye and R. W. England (eds), *Land value taxation. Theory, evidence, and practice*, Lincoln Institute of Land Policy, Cambridge MA, pp. 51–71.

Polzehl, J. and Spokoiny, V.: 2006, Propagation-separation approach for local likelihood estimation, *Probability Theory and Related Fields* **135**, 335–362.

Polzehl, J. and Spokoiny, V.: 2008, Structural adaptive smoothing by propagation-separation-methods, *in* C. Chen., W. K. Härdle and A. Unwin (eds), *Handbook of Data Visualization*, Springer, Berlin and Heidelberg, pp. 471–492.

Schulz, R., Wersing, M. and Werwatz, A.: 2014, Automated valuation modelling: A specification exercise, *Journal of Property Research* **31**, 131–153.

Senatorin für Finanzen, F. H. B.: 2010, Grundsteuer auf der Basis von Verkehrswerten, *Machbarkeitsstudie*, Bremen.

Sjoquist, D. L. and Walker, M. B.: 1999, Economies of scale in property tax assessment, *National Tax Journal* **52**, 207–220.

Vidakovic, B.: 1999, *Statistical Modeling by Wavelets*, first edn, Wiley, New York.

Watson, G. S.: 1964, Smooth regression analysis, *Sankhya: The Indian Journal of Statistics* **26**, 359–372.

Yatchew, A.: 1997, An elementary estimator of the partial linear model, *Economics Letters* **57**, 135–143.